How-To: Run a Cleanse Job

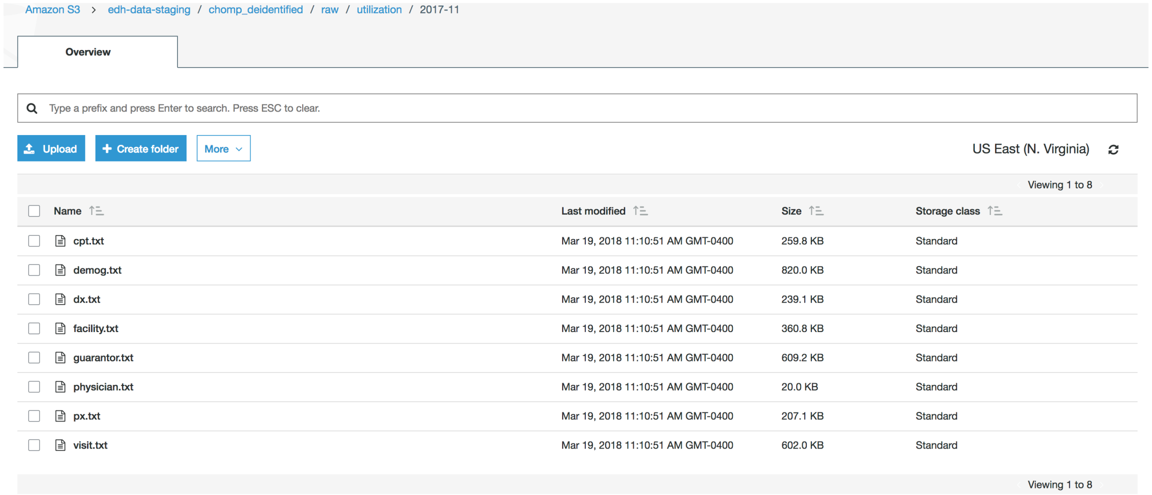
Primary data source providers for EDH

1. Client
    a. Call Center
    b. DNS (Do not Solicit)
    c. Persuade output
    d. Encounter
    e. Marketing lists
    f. Employee roster
    g. Physician roster
2. Experian
    a. New Movers
    b. Prospect
3. Social Security Administration
    a. Death Master File

Cleanse Utilization Job:

Reads utilization data for a customer and produces a .parquet file of cleansed activities, an error .csv file for dirty data and archives the processed utilization files.

Requires a S3 input bucket keys path:



Amazon S3 › edh-data-staging / chomp_deidentified / raw / utilization / 2017-11

Name	Last modified	Size	Storage class
cpt.txt	Mar 19, 2018 11:10:51 AM GMT-0400	259.8 KB	Standard
demog.txt	Mar 19, 2018 11:10:51 AM GMT-0400	820.0 KB	Standard
dx.txt	Mar 19, 2018 11:10:51 AM GMT-0400	239.1 KB	Standard
facility.txt	Mar 19, 2018 11:10:51 AM GMT-0400	360.8 KB	Standard
guarantor.txt	Mar 19, 2018 11:10:51 AM GMT-0400	609.2 KB	Standard
physician.txt	Mar 19, 2018 11:10:51 AM GMT-0400	20.0 KB	Standard
px.txt	Mar 19, 2018 11:10:51 AM GMT-0400	207.1 KB	Standard
visit.txt	Mar 19, 2018 11:10:51 AM GMT-0400	602.0 KB	Standard

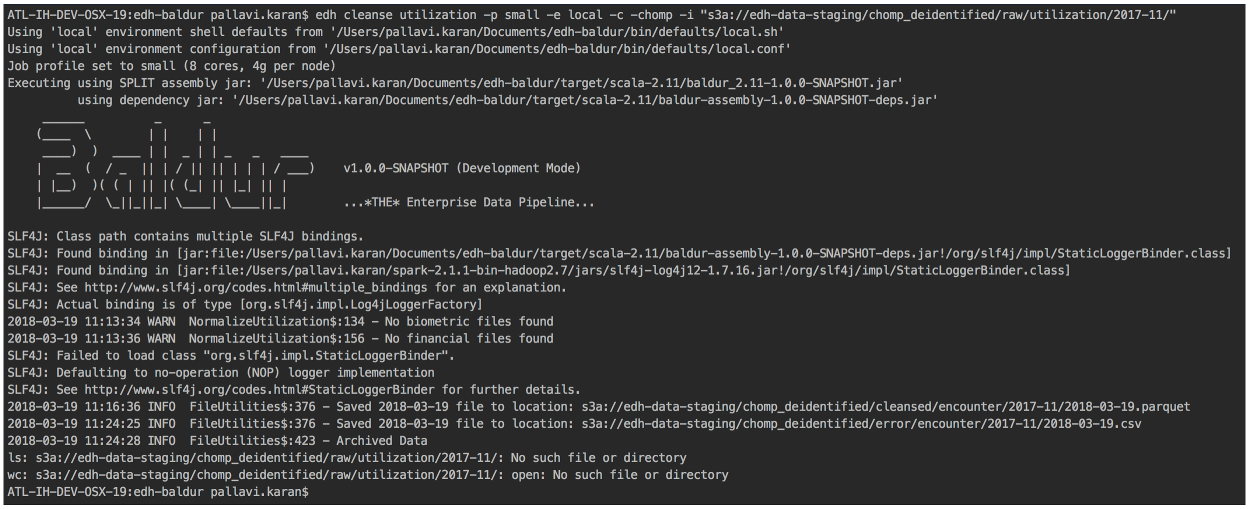
Command to Run the job:

```
(local) edh cleanse encounter -e local -p small -i
s3://<bucket-name>/<client>/raw/<activityType>/<format>/<batch>/
(qa) edh cleanse encounter -e qa -p med -i s3://<bucket-name>/<client>/raw/<activityType>/<format>/<batch>/

(stage) edh cleanse encounter -e stage -p large -i
s3://<bucket-name>/<client>/raw/<activityType>/<format>/<batch>/

(prod)  edh cleanse encounter -e prod -p xlarge -i
s3://<bucket-name>/<client>/raw/<activityType>/<format>/<batch>/
```

*Note: **-i** Specifies the input S3 folder that defines the location of incoming files*

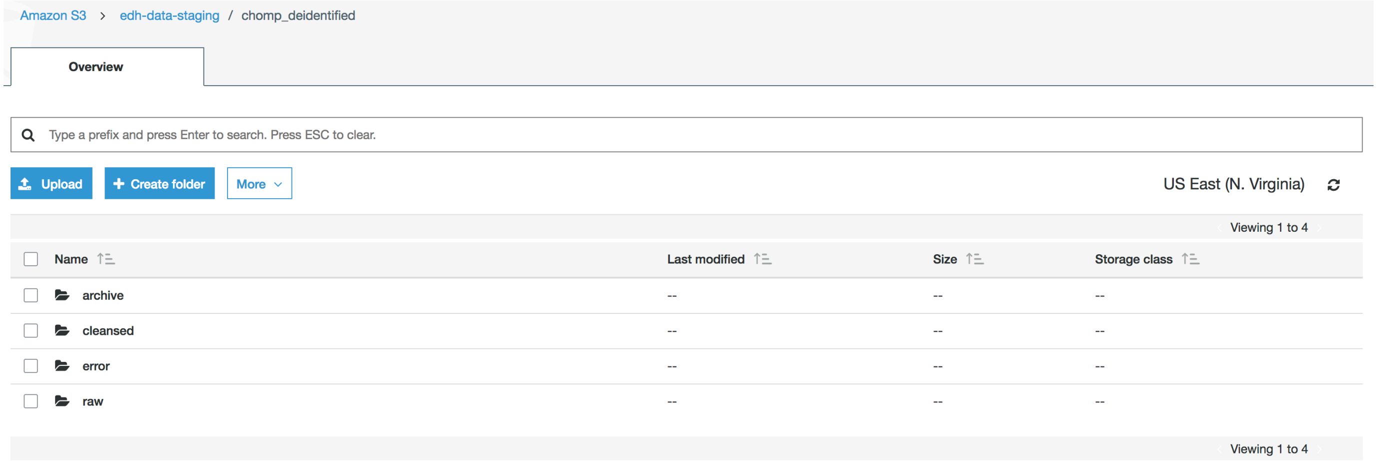
Example:

```
edh cleanse encounter -p small -e local -c chomp -i “s3a://edh-data-staging/chomp/raw/encounter/influencehealth/2017-09/” -f influencehealth
```



```
ATL-IH-DEV-OSX-19:edh-baldur pallavi.karan$ edh cleanse utilization -p small -e local -c -chomp -i "s3a://edh-data-staging/chomp_deidentified/raw/utilization/2017-11/"
Using 'local' environment shell defaults from '/Users/pallavi.karan/Documents/edh-baldur/bin/defaults/local.sh'
Using 'local' environment configuration from '/Users/pallavi.karan/Documents/edh-baldur/bin/defaults/local.conf'
Job profile set to small (8 cores, 4g per node)
Executing using SPLIT assembly jar: '/Users/pallavi.karan/Documents/edh-baldur/target/scala-2.11/baldur_2.11-1.0.0-SNAPSHOT.jar'
          using dependency jar: '/Users/pallavi.karan/Documents/edh-baldur/target/scala-2.11/baldur-assembly-1.0.0-SNAPSHOT-deps.jar'
     ______       _      _
    (____  \     | |    | |
     ____)  ) ____| |  _ | |_   _  ____
    |  __  ( / _  || | / || || | | |/ ___)    v1.0.0-SNAPSHOT (Development Mode)
    | |__)  )( ( | || |( (_| || |_| || |
    |______/  \_||_||_| \____| \____||_|        ...*THE* Enterprise Data Pipeline...

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/Users/pallavi.karan/Documents/edh-baldur/target/scala-2.11/baldur-assembly-1.0.0-SNAPSHOT-deps.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/Users/pallavi.karan/spark-2.1.1-bin-hadoop2.7/jars/slf4j-log4j12-1.7.16.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-03-19 11:13:34 WARN  NormalizeUtilization$:134 - No biometric files found
2018-03-19 11:13:36 WARN  NormalizeUtilization$:156 - No financial files found
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
2018-03-19 11:16:36 INFO  FileUtilities$:376 - Saved 2018-03-19 file to location: s3a://edh-data-staging/chomp_deidentified/cleansed/encounter/2017-11/2018-03-19.parquet
2018-03-19 11:24:25 INFO  FileUtilities$:376 - Saved 2018-03-19 file to location: s3a://edh-data-staging/chomp_deidentified/error/encounter/2017-11/2018-03-19.csv
2018-03-19 11:24:28 INFO  FileUtilities$:423 - Archived Data
ls: s3a://edh-data-staging/chomp_deidentified/raw/utilization/2017-11/: No such file or directory
wc: s3a://edh-data-staging/chomp_deidentified/raw/utilization/2017-11/: open: No such file or directory
ATL-IH-DEV-OSX-19:edh-baldur pallavi.karan$
```

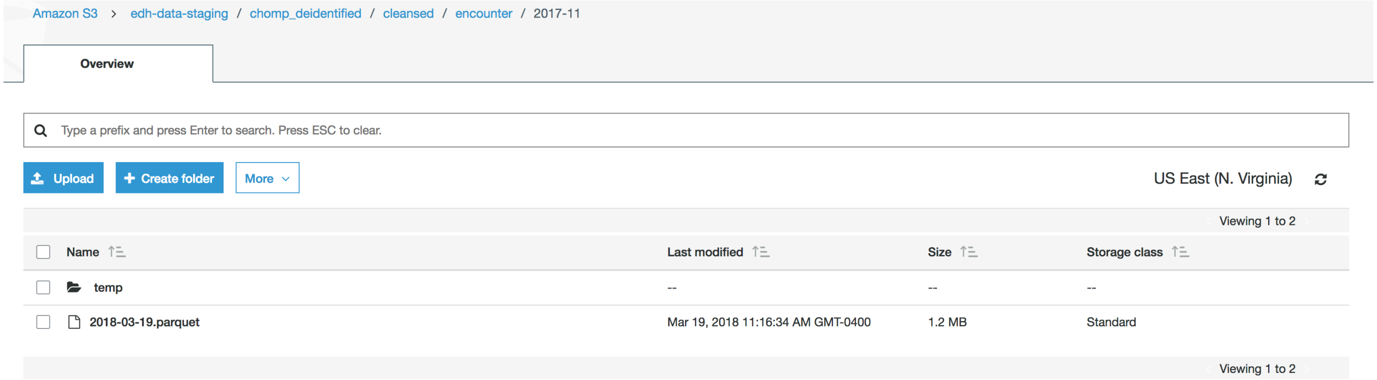
The outputs generated from the above job are saved in the following S3 bucket keys:



Amazon S3 › edh-data-staging / chomp_deidentified
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US East (N. Virginia)
Viewing 1 to 4

□ Name	Last modified	Size	Storage class
□ archive	--	--	--
□ cleansed	--	--	--
□ error	--	--	--
□ raw	--	--	--

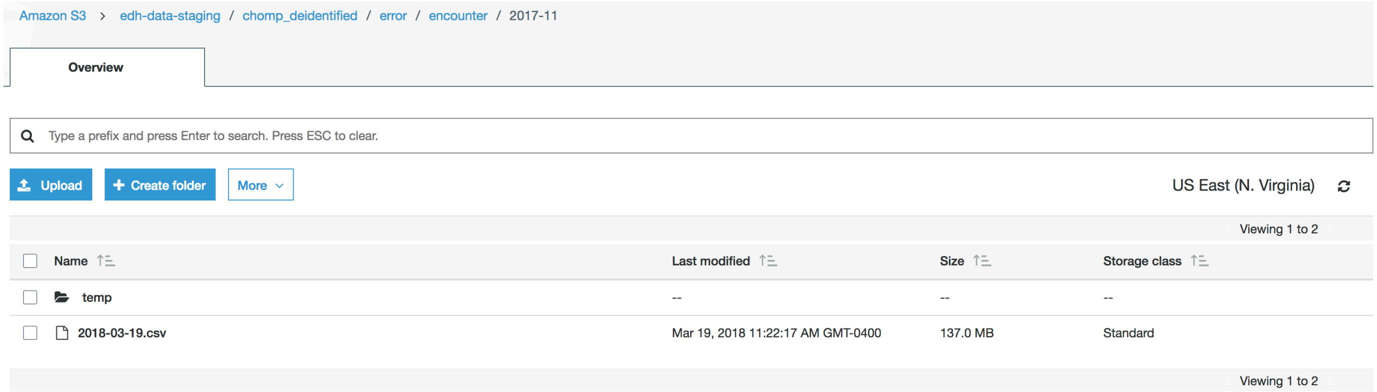
```
Example: s3a://edh-data-staging/chomp_deidentified/cleansed/encounter/2017-11/2018-03-16.parquet
```



Amazon S3 › edh-data-staging / chomp_deidentified / cleansed / encounter / 2017-11
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US East (N. Virginia)
Viewing 1 to 2

□ Name	Last modified	Size	Storage class
□ temp	--	--	--
□ 2018-03-19.parquet	Mar 19, 2018 11:16:34 AM GMT-0400	1.2 MB	Standard

```
Example: s3a://edh-data-staging/chomp_deidentified/error/encounter/2017-11/influencehealth/2018-03-16.csv
```



Amazon S3 › edh-data-staging / chomp_deidentified / error / encounter / 2017-11

□	Name	Last modified	Size	Storage class
□	temp	--	--	--
□	2018-03-19.csv	Mar 19, 2018 11:22:17 AM GMT-0400	137.0 MB	Standard

```
Example:
s3a://edh-data-staging/chomp_deidentified/archive/encounter/influencehealth/2017-11/chomp_deidentified-encounter-influencehealth-2017-11_2017-12-27T10-45-34.done
```